Exome-chip association analysis reveals an Asian-specific missense variant in PAX4 associated with type 2 diabetes in Chinese

Cheung, Chloe YY; Thomas, G Neil; Cheng, Kar

DOI:
10.1007/s00125-016-4132-z

Document Version
Peer reviewed version

Citation for published version (Harvard):
Cheung, CYY, Thomas, GN & Cheng, K 2016, 'Exome-chip association analysis reveals an Asian-specific missense variant in PAX4 associated with type 2 diabetes in Chinese', *Diabetologia*.
https://doi.org/10.1007/s00125-016-4132-z

Exome-chip association analysis reveals an Asian-specific missense variant in PAX4 associated with type 2 diabetes in Chinese

Thomas, G Neil; Cheng, Kar

Diabetologia Springer

Exome-chip association analysis reveals an Asian-specific missense variant in PAX4 associated with type 2 diabetes in Chinese

Journal:	*Diabetologia*
Manuscript ID	Diab-16-0753.R2
Manuscript Type:	Article
Keywords:	3.01.03 Genetics of type 2 diabetes, 3.01 Genetics / Epidemiology (all), 2.10 Human

Exome-chip association analysis reveals an Asian-specific missense variant in *PAX4* associated with type 2 diabetes in Chinese

Chloe YY Cheung[1]*, Clara S Tang[2]*, Aimin Xu[3,4,5], Chi-Ho Lee[1], Ka-Wing Au[1], Lin Xu[6], Carol HY Fong[1], Kelvin HM Kwok[1], Wing-Sun Chow[1], Yu-Cho Woo[1], Michele MA Yuen[1], JoJo SH Hai[1], Ya-Li Jin[7], Bernard MY Cheung[1], Kathryn CB Tan[1], Stacey S Cherny[8], Feng Zhu[7], Tong Zhu[7], G.Neil Thomas[9], Kar-Keung Cheng[9], Chao-Qiang Jiang[7], Tai-Hing Lam[6,7]**, Hung-Fat Tse[1,10]**, Pak-Chung Sham[8,11,12]**, Karen SL Lam[1,3,4]**

[1]Department of Medicine, the University of Hong Kong, Hong Kong, China; [2]Department of Surgery, the University of Hong Kong, Hong Kong, China; [3]State Key Laboratory of Pharmaceutical Biotechnology, The University of Hong Kong, Hong Kong, China; [4]Research Centre of Heart, Brain, Hormone and Healthy Aging, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China; [5]Department of Pharmacology & Pharmacy, The University of Hong Kong, Hong Kong, China; [6]School of Public Health, the University of Hong Kong, Hong Kong, China; [7]Guangzhou No.12 Hospital, Guangzhou 510620, China; [8]Department of Psychiatry, the University of Hong Kong, Hong Kong, China; [9]Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom; [10]Hong Kong-Guangdong Joint Laboratory on Stem Cell and Regenerative Medicine, the University of Hong Kong, Hong Kong, China; [11]Centre for Genomic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China; [12]State Key Laboratory in Brain and Cognitive Sciences, The University of Hong Kong, Hong Kong, China.

*CYY Cheung and CS Tang contributed equally to this work and should be considered as co-first authors; **KSL Lam, PC Sham, HF Tse and TH Lam contributed equally to the supervision of this work and are co-corresponding authors.

Correspondence:

Karen SL Lam: Department of Medicine, The University of Hong Kong, Queen Mary Hospital, 102 Pokfulam Road, Hong Kong, China; Phone: +852-2255-3348/ +852-2255-4783; Fax: +852-2816-2863; Email: ksllam@hku.hk

Pak-Chung Sham: Centre for Genomic Sciences, The University of Hong Kong, 6/F, HKJC Building for Interdisciplinary Research, 5 Sassoon Road, Pokfulam Hong Kong, China; Phone: +852-2831-5425; Email: pcsham@hku.hk

Hung-Fat Tse: Department of Medicine, The University of Hong Kong, Rm 1928, Block K, Queen Mary Hospital, 102 Pokfulam Road, Hong Kong, China; Phone: +852-2831-4694; Email: hftse@hkucc.hku.hk

Tai-Hing Lam: School of Public Health, Rm 505, Faculty of Medicine Building, William M.W. Mong Block, The University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong, China; Phone: +852-3917-9287; Email: hrmrlth@hku.hk

Word count (main text): 3932; **Word count (abstract):** 250

Abstract (Word count:250)

Aims:

Genome-wide association studies (GWAS) identified many common type 2 diabetes-associated variants, mostly found at the intronic or intergenic regions. Recent advancement of exome array genotyping platforms have opened up a novel means for detecting the associations of low-frequency or rare coding variants with type 2 diabetes. We conducted an exome-chip association analysis to identify additional type 2 diabetes susceptibility variants in the Chinese population.

Methods:

An exome-chip association study was conducted by genotyping 5640 Hong Kong Chinese individuals using a custom Asian Exome-chip. Single variant association analysis was conducted on 77,468 single nucleotide polymorphisms (SNPs). Fifteen SNPs were subsequently genotyped for replication analysis in an independent Chinese cohort comprising 12,362 Guangzhou individuals. A combined analysis involving 7189 cases and 10,813 controls was performed.

Results:

In the discovery stage, an Asian-specific coding variant rs2233580 (p.Arg192His) in *PAX4*, and 2 variants at the known loci, *CDKN2B-AS1* and *KCNQ1,* were significantly associated with type 2 diabetes at exome-wide significance ($p_{discovery}<6.45x10^{-7}$). The risk allele (T) of *PAX4* rs2233580 was associated with a younger age of diabetes diagnosis. This variant was replicated in an independent cohort and demonstrated a stronger association that reached genome-wide significance ($p_{meta=}3.74x10^{-15}$) in the combined analysis.

Conclusions:

We identified the association of a *PAX4* Asian-specific missense variant rs2233580 with type 2 diabetes in an exome-chip association analysis, supporting the involvement of *PAX4* in the

pathogenesis of type 2 diabetes. Our findings are suggestive of *PAX4* being a possible effector gene of the 7q32 locus previously identified from GWAS amongst Asians.

Key words: Asian-specific; Exome-chip association analysis; *PAX4;* Type 2 diabetes

Abbreviations:

CRISPS	Hong Kong Cardiovascular Risk Factor Prevalence Study
CSN1S1	Casein Alpha S1
FGF21	Fibroblast growth factor 21
FPG	Fasting plasma glucose
GBCS	Guangzhou Biobank Cohort Study
GWAS	Genome-wide association studies
HKU-TRS	The University of Hong Kong Theme-based Research Scheme
HKWDR	Hong Kong West Diabetes Registry
MAF	Minor allele frequency
MODY	Maturity-onset diabetes of the young
PAX4	Paired Box Gene 4
PC	Principal component
SNP	Single nucleotide polymorphism
TTBK2	Tau-tubulin-kinase 2

Introduction

Type 2 diabetes is a common disease resulting from the complex interactions between multiple genetic and environmental factors. Insights into the genetic basis of type 2 diabetes will facilitate the discovery of novel treatment targets. Since 2007, the success in genome-wide association studies (GWAS) has led to the identification of a large number of independent loci for type 2 diabetes. However, the disease-susceptibility single nucleotide polymorphisms (SNPs) identified from these GWAS are common variants which tend to confer relatively small effect sizes, altogether accounting for only 10 to 15% of the type 2 diabetes heritability [1]. The functional consequences of these susceptibility variants, which are mostly present in intronic or intergenic regions, remained difficult to interpret. In the past few years, the role of low-frequency (minor allele frequency [MAF] =1%-5%) and rare (MAF<1%) coding variants with various complex traits [2-7] is being increasingly studied. It has been suggested that most current rare variants were introduced by mutational events during the recent explosive growth of the human population [8, 9]. These rare variants are believed to confer a greater effect than the common variants due to the limited time for purifying selection to act [9, 10]. The majority of efforts to reveal type 2 diabetes susceptibility variants have been made in populations of European ancestry. Utilising the advanced technologies, such as the exome-chip and whole genome/exome-sequencing, researchers have detected associations of additional novel coding variants, both common and rare, for type 2 diabetes [4, 5] and several quantitative glycaemic traits, such as fasting glucose and insulin levels in European populations [3, 6, 7]. As samples of European ancestry represent only a subset of human genetic variations [11], the risk variants in other populations are likely to be insufficiently characterised. A genome-wide trans-ancestry meta-analysis reported several type 2 diabetes-susceptibility variants which showed significant difference in effect sizes and associations in different populations [12]. For instances, the effect size of *TCF7L2* rs7903146 was higher in Europeans than in East Asians; and the association signal of *PEPD* rs3786897 was specific to the populations of East Asians, while the association signal of *KLF14* rs13233731 was only significant in the European samples [12]. Such observations highlight the importance of conducting

association analysis in non-European populations to detect novel loci affecting the risk of type 2 diabetes.

The advancement in array-based genotyping technology, such as exome arrays, has provided a more cost effective approach, compared to whole-genome or exome sequencing, for assessing the association of rare and low-frequency coding variants which may be population-specific. In a joint collaboration study, our group has recently reported several novel or Asian-specific coding variants associated with blood lipids [2], using a tailored Illumina HumanExome BeadChip (Asian Exome-chip [13]). In the present study, we aimed to detect novel loci for type 2 diabetes in the Chinese population using the Asian Exome-chip. We first conducted an exome-chip association analysis based on 5640 participants from the University of Hong Kong Theme-based Research Scheme (HKU-TRS) cohort, and genotyped 15 SNPs for replication in an independent Southern Han Chinese cohort from Guangzhou (n=12,362).

Methods and Materials

Participants

Discovery cohort:

The discovery stage involved a total of 5640 Southern Han Chinese participants (3652 cases and 1988 controls) from the HKU-TRS cohort who participated in a previous exome-chip association study for blood lipid traits [2]. The study participants were recruited from the Hong Kong West Diabetes Registry (HKWDR) [14]; the Hong Kong Cardiovascular Risk Factor Prevalence Study (CRISPS) [15] and the Chinese CAD cohort of the Queen Mary Hospital in Hong Kong. Details of the corresponding cohorts have been previously reported [2]. Type 2 diabetes cases were defined as fulfilling at least one of the following criteria: fasting plasma glucose (FPG)≥7mmol/l; or 2-hr glucose during oral glucose tolerance test (OGTT)≥11.1mmol/l; or taking glucose-lowering agents; or physician-diagnosed diabetes. All controls had no documented history of diabetes and were not on treatment for diabetes. Written informed consent was obtained from each participant and the study protocol was approved by

the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster.

Replication cohort:

The replication stage involved a total of 12,362 Southern Han Chinese participants (3537 cases and 8825 controls) of the Guangzhou Biobank Cohort Study (GBCS) (ESM Table 1). The clinical characteristics and glycaemic status were based on cross-sectional data obtained at the time of blood sample collection. Details of GBCS have been described previously [16]. In brief, GBCS is a collaborative project between Guangzhou Number 12 Hospital, the University of Hong Kong and the University of Birmingham. The GBCS was established to examine the effect of genetic and environmental influences on health problems and development of chronic diseases. The baseline recruitment was conducted from 2003 to 2008 (n=30,519; aged 50 or above) in Guangzhou [17]. Participants were invited to participate in the second examination from August 2008 to December 2012. The present study included participants who attended the second examination and had sufficient information for determination of type 2 diabetes status. Type 2 diabetes cases were defined as fulfilling at least one of the following criteria: FPG≥7mmol/l; or 2-hr glucose during OGTT≥11.1mmol/l; or haemoglobin A1c (HbA1c) ≥6.5% (47.5mmol/mol); or taking glucose-lowering agents; or self-reported physician-diagnosed diabetes. All controls were not on treatment for diabetes, had no documented history of diabetes and had FPG <6.1mmol/l and 2-hr glucose during OGTT<7.8mmol/l. Written informed consent was obtained from each participant and the study protocol was approved by the Guangzhou Medical Ethics Committee of the Chinese Medical Association.

Genotyping and data quality control

Discovery stage:

All participants were genotyped using a custom Asian Exome-chip which was a specially designed exome array with an add-on content of 58,317 variants in addition to the standard content of the

Infinium HumanExome BeadChip (HumanExome-12v1_A; Illumina, CA). Detailed description of the Asian Exomechip design has been presented elsewhere [2, 13]. Briefly, the standard content of the exome array includes 242,901 markers, including>200,000 protein-altering variants identified from approximately 12,000 sequenced genomes and exomes of primarily European ancestry; and >20,000 non-exonic variants contributed by multiple consortia, such as NHLBI Exome Sequencing Project; and variants designed for ancestry differentiation, sample tracking and for establishing segments of identity by descent (http://genome-sph.umich.edu/wiki/Exome_Chip_Deisgn; accessed 1 May 2016). The European based design has led to an under-representation of the non-European genomes and thereby limited the coverage of low frequency variants among the non-European populations. With a view to allow comprehensive genotyping across the full allele frequency spectrum in Asians, a custom panel of ~30,000 missense or nonsense variants identified from 3 independent Asian sequencing data sets of ~1,000 Chinese samples were integrated into the Asian Exomechip. Additionally, a custom set of common variants selected for GWAS follow-up or fine mapping studies was also included. Genotype calling was conducted by the GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina). We first conducted manual inspection of genotype clusters for over 55,000 variants that had a GenTrain score <0.8; or with high missingness (>1%); or were shown to have poor genotype clustering in exome chip genotyping of >9000 individuals by collaborators [13, 18]. A total of 4550 variants with poor genotype clustering were removed. Individual-level quality control (QC) was conducted with regard to gender mismatch, duplication, biological relatedness, and possible sample contamination. A principal component (PC) analysis was conducted to examine for the existence of non-Chinese samples using a panel of >20,000 independent common SNPs (MAF>0.05), with outliers excluded from the analysis. For SNP-level QC, we excluded 217,455 SNPs with MAF<0.1%, of which 179,107 SNPs are monomorphic; 154 SNPs which deviated from Hardy-Weinberg Equilibrium (HWE) with $p<1x10^{-5}$ in controls; 3854 SNPs with >2% missingness; and 8,086 SNPs that were originally designed for the purpose of QC, including the fingerprint SNPs for sample tracking, ancestry informative markers (AIMs) for distinguishing Europeans from native and African

Americans, and grid SNPs for the identification of identity by descent segments. After QC measures, a total of 5640 participants and 77,468 variants were included in the association analysis.

Replication stage:

In the replication stage, we genotyped all SNPs which achieved $p_{discovery}<5x10^{-4}$ and with potential functional relevance, except *CDKN2B-AS1/DMRTA1* rs10965250 and *KCNQ1* rs2237896, which had been previously reported to be of genome-wide significance ($p<5x10^{-8}$) in GWAS [1], and also reached exome-wide significance ($p_{discovery}<6.45x10^{-7}$ [=0.05/77,468]) in the current study. By SNPs with potential functional relevance, we refer to SNPs at or near genes that showed protein-protein interactions, or shared same pathway with known type 2 diabetes susceptibility genes, or were implicated in the pathogenesis of diabetes. These included *PAX4* rs2233580, *CDKAL1* rs10440833, *FGFR1* rs2288696, *ANKRD55/MAP3K1* rs456867, *IGF2BP2* rs11711477, *TTBK2* rs56017612 and *DUSP26/UNC5D* rs4739563, *HCG27/HLA-C* rs3869115, *SCN1B* rs67701503, *DAP* rs267939, *CSN1S1* rs10030475, *ZNF283/ZNF404* rs138993781, *STAB1* rs740903, *CARNS1* rs868167, and *PDPN* chr1:13937002. All 15 selected SNPs were then genotyped using the MassARRAY Sequenom platform (San Diego, CA, USA) at the Beijing Genomics Institute (BGI), Beijing. Four SNPs which either showed low genotyping call rate (<90%; *PDPN* chr1:13937002 and *IGF2BP2* rs11711477); or deviated from HWE in controls ($p_{HWE}<0.003=0.05/15$; *CDKAL1* rs10440833 and *SCN1B* rs67701503) in the replication study were excluded from further analysis. Thus for the final analysis, a total of 11 SNPs were included. *PAX4* rs2233580 did not deviate from HWE in controls and was therefore retained in the analysis, even though it was significantly deviated from HWE in the case group ($P_{HWE}<0.003$). *PAX4* has been implicated in the pathogenesis of type 2 diabetes [19], and it is recognised that a true association can lead to deviation from HWE in cases [20]. The average genotyping call rate of these SNPs was 98.2%.

Data analysis

All statistical analyses in the discovery and replication stages were conducted using PLINK version 1.9 [21]. In the discovery stage, multiple logistic regression analysis with adjustment for age, sex and first 2 PCs was employed to examine for the associations with type 2 diabetes, under the additive genetic model. To assess the adiposity independent association of the top SNPs with type 2 diabetes, we further included body mass index (BMI) in the multiple logistic regression model. Exome-wide significance was defined as $p<6.45\text{x}10^{-7}$ (=0.05/77,468). To address the between-SNP linkage disequilibrium (LD), the *p*-value informed LD based clumping approach with the "--clump" command implemented in PLINK was conducted. The index SNP had the most significant *p*-value from each clumped association region. Each index SNP formed clumps with other variants which were in LD with the index SNP ($r^2 \geq 0.2$) and were within ± 500kb from the index SNP. The association between *PAX4* rs2233580 and age of diabetes diagnosis was examined by univariate linear regression analysis. In the replication stage, age and sex were included as covariates in the multiple logistic regression model to assess for associations with type 2 diabetes. A Bonferroni corrected one-tailed *p*-value $<4.54\text{x}10^{-3}$ (=0.05/11) was used as the threshold for successful replication. Meta-analysis of the association results of the discovery and replication stages were conducted using METAL [22]. Inverse variance fixed-effect method was employed to pool the summary statistics of the two stages and heterogeneity of effect was assessed using Cochran's Q-test and I^2 index.

Variants annotation and *in silico* functional analysis

Function of variants and protein changes for non-synonymous SNPs were annotated by KGGSeq [23] according to the RefGene annotation. The pathogenic potential of the non-synonymous variants were assessed through various deleteriousness and conservation prediction tools implemented in KGGseq, including SIFT [24] and PolyPhen [25].

Asian-specific variants

Variants were classified as "Asian-specific" if they were monomorphic in both the European and African populations but polymorphic (MAF>0) in the Asian population, according to the 1000 Genomes Project [11].

Results

A total of 5640 Chinese (Hong Kong) participants were genotyped using a custom Asian Exomechip (Table 1). Single variant association analysis was performed to assess the associations with type 2 diabetes for 77,468 polymorphic variants (Figure 1). Of these, 48% alter protein composition and 21% were Asian-specific variants with MAF between 0.1 to 5%.

In the discovery stage, single variant association analysis was conducted in 3652 cases and 1988 controls, adjusted for age, sex and the first 2 PCs. We detected 34 index SNPs within 32 loci significantly associated with type 2 diabetes at $p_{discovery}<5x10^{-4}$ (ESM Table 2), of which, 3 variants reached exome-wide significance ($p_{discovery}<6.45x10^{-7}$) (Table 2). These included the known associations at *CDKN2B-AS1/DMRTA1* rs10965250 ($p_{discovery}=5.93x10^{-8}$, OR[95%CI]:0.80[0.74, 0.87]) and *KCNQ1* rs2237896 ($p_{discovery}=1.82x10^{-7}$; OR[95%CI]:0.80[0.73, 0.87]) reported in previous GWAS, as well as an Asian-specific variant, rs2233580 (p.Arg192His), of *PAX4* which showed an exome-wide significant association with type 2 diabetes ($p_{discovery}=1.75x10^{-7}$; OR[95%CI]:1.39[1.23, 1.56]). As *PAX4* is a known gene for maturity onset diabetes of the young (MODY) [26], we further examined its association with age of diabetes diagnosis. The risk allele (T) of rs2233580 was found to be significantly associated with younger age of diabetes diagnosis ($p=6.01x10^{-4}$; beta[95%CI]: -1.45[-2.28, -0.62]; Mean age of diagnosis ± standard deviation [years]: TT:52±13; CT:53±13; CC:54±13). In addition, we also identified several loci not previously reported to be associated with type 2 diabetes: *FGFR1* rs2288696 ($p_{discovery}=2.29x10^{-5}$; OR:[95%CI]:0.73[0.63, 0.85]), *TTBK2* rs56017612 ($p_{discovery}=7.40x10^{-5}$; OR[95%CI]:0.72[0.61, 0.85]) and *DUSP26/UNC5D* rs4739563

($p_{discovery}$=7.48x10^{-5}; OR:[95%CI]:0.80[0.0.72, 0.90]). Association of all SNPs remained significant after further adjustment for BMI (ESM Table 2).

In the replication stage, 11 of the 15 selected SNPs passed QC and were analysed in 3537 cases and 8825 controls. Replication and combined association results of these SNPs are shown in Table 3. Of these, 8 of them showed consistent direction of effects. Only the association of the *PAX4* missense variant rs2233580 with type 2 diabetes was successfully replicated (one-tailed $p_{replication}$=1.22x10^{-9}; OR[95%CI]:1.28[1.18, 1.39]; remained significant after Bonferroni correction). Meta-analysis of the association results gave a genome-wide significant association, with no evidence of heterogeneity in effect size (p_{meta}=3.74x10^{-15}, OR[95%CI]:1.31[1.23, 1.40]; I^2=10, $p_{heterogeneity}$=0.292). The associations of *FGFR1* rs2288696, *TTBK2* rs56017612 and *DUSP26/UNC5D* rs4739563 were not significant in the replication cohort. However, the direction of effects for both *FGFR1* rs2288696 and *TTBK2* rs56017612 were consistent with those from the discovery stage. A modest association was observed at a missense variant of *CSN1S1* (rs10030475 [p.Pro137Thr]; one-tailed $p_{replication}$=7.5x10^{-3}, OR[95%CI]:0.93[0.87, 0.99]). However, this association did not pass Bonferroni correction for multiple testing in the replication stage.

Conclusions

The present study reports the first exome-chip association analysis on type 2 diabetes in a Chinese population. By genotyping 5640 Chinese participants using a custom Asian Exome-chip which interrogated 77,468 polymorphic SNPs, we identified the association of an Asian-specific coding variant in *PAX4* and replicated the associations of some known type 2 diabetes-susceptibility loci. We also detected a few possible candidates which showed potential functional relevance in the pathogenesis of type 2 diabetes, such as *TTBK2, FGFR1* and *CSN1S1*.

The identification of the Asian-specific and probably damaging variant of *PAX4* is the major finding of this study. *PAX4* encodes a member of the PAX family of paired-homeodomain factor. *PAX4* functions as a transcription repressor and plays a crucial role in pancreatic beta-cell function and development [27]. It also plays a role in beta-cell proliferation and survival [28, 29]. The heterozygous *PAX4* knockout mice harbour less mature pancreatic beta- and delta-cells but with numerous abnormally clustered alpha-cells, indicating the essential role of *PAX4* in the differentiation of beta- and delta-cell lineages [30]. *PAX4* has been shown to repress transcriptional activity of insulin [19] and glucagon [31] promoters. *PAX4* is located at 7q32, a region reported to be associated with type 2 diabetes in previous GWAS of Asians [32, 33]. An intergenic variant rs6467136 located near *GRIP* and *GCC1-PAX4* was reported to be associated with type 2 diabetes in a meta-analysis of 8 GWAS studies in East Asians [32]; and rs10229583 located downstream of *PAX4* was identified in a GWAS for type 2 diabetes in a Chinese population [33]. Such observations, together with our findings, suggested that the effect of *PAX4* may be more evident in East Asians than in other populations. Both of these SNPs appear to be independent to the missense variant rs2233580 identified in the current study. rs2233580 shows very low LD with both rs6467136 (r^2=0.03) and rs10229583 (r^2=0.02). The association of rs6467136 was not significant ($p_{discovery}$=0.284). Data for rs10229583 was not available for analysis. Our findings have provided evidence that *PAX4* is a possible effector gene at 7q32, a GWAS locus for type 2 diabetes. Our exome-chip could achieve 50% coverage of the coding variants within this gene region. Nonetheless, we were unable to eliminate the possibility that the association of rs2233580 identified in the current study resulted from tagging of other causative coding variants which were not covered by our exome-chip. On the other hand, its functional significance, as demonstrated by *in silico* [24, 25] and *in vitro* [34, 35] analyses, suggests that this SNP is likely to be the causative variant. While *in silico* analysis of the 2 previously reported intergenic variants was unable to give a clue to their functional relevance (RegulomeDB score=5 for rs10229583; 6 for rs6467136), rs2233580 was predicted to be damaging by multiple prediction tools [24, 25] (SIFT score=0; PolypPhen2 HDIV score=1; Polyphen2 HVAR score=0.99). *In vitro* study showed that the transcriptional repressor activities of *PAX4* p.Arg192His on human insulin and glucagon promoters

were reduced when compared with the wild-type *PAX4* [34]. The Arg192 residue is highly conserved across different species, including human, mouse, rat and chimpanzee and this residue has been shown to make a direct contact with the major groove of the DNA binding sequence [35]. An amino acid change in the homeodomain of PAX4 may cause a defect in its transcriptional activity. It has been proposed that this variant may affect diabetes risk through its effect on beta-cell proliferation in adult pancreas, or beta-cell differentiation and maturation during development which leads to beta-cell mass reduction [34]. While rs2233580 has a frequency of ~10% among Asian populations, this variant was found to be monomorphic in Europeans and Africans, suggesting interrogation in less studied non-European populations would facilitate the identification of novel population-specific associations. Our finding of an Asian-specific variant also has implications on the construction of polygenic genetic scores to predict type 2 diabetes in Asian populations.

Our observation of the significant association of *PAX4* rs2233580 with type 2 diabetes was in agreement with findings from a large-scale whole-genome/exome sequencing study conducted by the GoT2D and T2D-GENES consortia [36], recently published during the review process of our manuscript. rs2233580 was reported to be associated with type 2 diabetes exclusively in 2165 East Asian individuals at genome-wide significance (p=9.3x10^{-9}), and this association was further replicated in 3 independent East Asian cohorts [36]. Mutations in *PAX4* have been found to cause the rare monogenic form MODY in Thais [26]. On the other hand, common variants of a number of established MODY genes have been found to be associated with type 2 diabetes, including *GCK*, *HNF1-aplha*, *HNF4-alpha*, *HNF1-beta*, and *PDX1* [37-39]. Findings from the current study and those reported by the GoT2D and T2D-GENES consortia suggest that *PAX4* also harbours common variants that confer susceptibility to type 2 diabetes. Interestingly, in a previous GWAS of East Asians, the risk allele of a common variant, rs10229583, located downstream of *PAX4,* was reported to be associated with higher risk of type 2 diabetes and a younger age of diagnosis [33]. Among the 3652 cases in the current study, individuals who carried the risk allele (T) of the *PAX4* missense variant rs2233580 were also significantly younger at the time of diagnosis. In contrast, the GoT2D and T2D-

GENES consortia reported no significant association between rs2233580 and age of diagnosis in a total of 1619 cases from the 3 independent cohorts of East Asian ancestry (Hong Kong Chinese, Korean, and Singapore Chinese) [36]. This contradictory observation could be attributable to the much larger sample size of the current study which provided sufficient power to detect the association (ESM Table 3). Furthermore, study heterogeneity caused by different ascertainment criteria of type 2 diabetes cases between studies may have also contributed to the discordant observations. A meta-analysis of our data with those of the 3 independent cohorts has provided evidence to support the association of *PAX4* rs2233580 with younger age of diagnosis (p_{meta}=0.007; z-score= -2.717; I^2=58.5, $p_{heterogeneity}$=0.065) (ESM Table 3).

Although unable to reach genome/exome-wide significance, the potential functions of *TTBK2*, *FGFR1* and *CSN1S1* have made them possible candidates for T2DM. TTBK2 (Tau-tubulin-kinase 2) is a serine/threonine kinase known to phosphorylate tau and tubulin [40]. TTBK2 is involved in the regulation of a sodium-dependent glucose transporter, SGLT1 [41], which is responsible for the absorption of glucose and galactose in the intestine and is involved in the renal reabsorption of glucose in kidney [42]. Depletion of TTBK2 has been shown to decrease SGLT1 stability in the cell membrane and lead to loss of glucose transport capacity in *Xenopus* oocytes [41]. Mice with attenuated FGFR1 signalling exhibited a reduced number of beta-cells, impaired expression of glucose transporter 2, enhanced proinsulin content in beta-cells and developed diabetes with age [43]. FGFR1 is the primary receptor of fibroblast growth factor 21 (FGF21), and hence regulates FGF21 responsiveness. FGF21 has shown beneficial metabolic effects in animals and humans [44]. Our team previously demonstrated that high FGF21 levels could predict type 2 diabetes development [15]. The paradoxical increase in FGF21 levels in patients with type 2 diabetes suggest that FGF21 resistance may play a role in the pathogenesis of type 2 diabetes [44]. Our finding that a variant of *FGFR1* is associated with type 2 diabetes is supportive of such a possibility. Casein Alpha S1 (*CSN1S1*) is a member of the casein family. CSN1S1 has been shown to possess proinflammatory properties such as the upregulation of IL-1beta [45], the causative role of which in the loss of beta-cell mass in type 2

diabetes has been suggested by data from animal studies and clinical trials [46]. Given their potential functional relevance in the pathogenesis of type 2 diabetes, more detailed investigation of these genes, such as deep sequencing analysis, is warranted.

A potential limitation of the present study was the under-representation of rare functional variants specific to the Chinese populations in the exome-chip. With an attempt to ameliorate this limitation, we included additional coding variants to augment the coverage. Small sample size has always been a major limitation hindering the identification of rare variants, as demonstrated in this study. The sample size of the discovery stage is relatively small and therefore lacks statistical power to detect variants with modest effect size or very low frequency. Future large-scale meta-analysis with other Asian cohorts may serve to identify more functional variants that are specific to our population. Trans-ethnic meta-analysis will help to enhance the fine-mapping resolution of causal variants. Another limitation of the present study could be the SNP selection for replication.

In summary, the significant association of an Asian-specific coding variant rs2233580 (p.Arg192His) with type 2 diabetes was identified in an exome-chip association analysis in a Chinese population. Our findings has provided compelling evidence that *PAX4* could be a possible effector gene of the 7q32 locus and supported its involvement in the pathogenesis of type 2 diabetes.

Acknowledgements

The authors thank all the study participants, clinical and research staffs of HKU-TRS and GBCS for their contribution in this research study.

Funding

This work was supported by the Hong Kong Research Grant Council: Theme Based Research Scheme (T12-705/11) and Collaborative Research Fund (HKU2/CRF/12R); The University of Hong Kong Foundation for Educational Development and Research (SN/1f/HKUF-DC;C20400.28505200); the

Guangzhou Public Health Bureau (201102A211004011); and the Guangzhou Science and Technology Bureau, Guangzhou, China (2002Z2-E2051; 2012J5100041; 2013J4100031).

Duality of interest: The authors declare that there is no duality of interest associated with this manuscript**.**

Contribution statement

KSLL, PCS, HFT and THL conceived the study, undertook project leadership and are guarantors of this work. CYYC and CST analysed the data and wrote the first draft of the manuscript. All authors contributed to the drafting and critical revision of the manuscript. PCS, CST and SSC provided useful comments to data-analysis. CYYC, AX, CHL, KWA, LX, CHYF, KHMK, WSC, YCW, MMAY, JSHH, YLJ, BMYC, KCBT, FZ and TZ were involved in the sample collection, selection and phenotype data preparation for the HKU-TRS and GBCS cohorts. KSLL, HFT, THL, GNT, KKC, CQJ were involved in the database management for the HKU-TRS and GBCS cohorts. All authors approved the final version of the manuscript.

References

[1] McCarthy MI (2010) Genomics, type 2 diabetes, and obesity. N Engl J Med 363: 2339-2350

[2] Tang CS, Zhang H, Cheung CY, et al. (2015) Exome-wide association analysis reveals novel coding sequence variants associated with lipid traits in Chinese. Nat Commun 6: 10206

[3] Huyghe JR, Jackson AU, Fogarty MP, et al. (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45: 197-201

[4] Albrechtsen A, Grarup N, Li Y, et al. (2013) Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. Diabetologia 56: 298-310

[5] Steinthorsdottir V, Thorleifsson G, Sulem P, et al. (2014) Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet 46: 294-298

[6] Wessel J, Chu AY, Willems SM, et al. (2015) Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat Commun 6: 5897

[7] Mahajan A, Sim X, Ng HJ, et al. (2015) Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define an effector transcript at the G6PC2-ABCB11 locus. PLoS Genet 11: e1004876

[8] Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740-743

[9] Nelson MR, Wegmann D, Ehm MG, et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337: 100-104

[10] Tennessen JA, Bigham AW, O'Connor TD, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64-69

[11] Abecasis GR, Auton A, Brooks LD, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65

[12] Mahajan A, Go MJ, Zhang W, et al. (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 46: 234-244

[13] Zhang Y, Long J, Lu W, et al. (2014) Rare coding variants and breast cancer risk: evaluation of susceptibility Loci identified in genome-wide association studies. Cancer Epidemiol Biomarkers Prev 23: 622-628

[14] Hui E, Yeung CY, Lee PC, et al. (2014) Elevated circulating pigment epithelium-derived factor predicts the progression of diabetic nephropathy in patients with type 2 diabetes. J Clin Endocrinol Metab 99: E2169-2177

[15] Chen C, Cheung BM, Tso AW, et al. (2011) High plasma level of fibroblast growth factor 21 is an Independent predictor of type 2 diabetes: a 5.4-year population-based prospective study in Chinese subjects. Diabetes Care 34: 2113-2115

[16] Jiang C, Thomas GN, Lam TH, et al. (2006) Cohort profile: The Guangzhou Biobank Cohort Study, a Guangzhou-Hong Kong-Birmingham collaboration. Int J Epidemiol 35: 844-852

[17] Jiang CQ, Lam TH, Lin JM, et al. (2010) An overview of the Guangzhou biobank cohort study-cardiovascular disease subcohort (GBCS-CVD): a platform for multidisciplinary collaboration. J Hum Hypertens 24: 139-150

[18] Guo Y, He J, Zhao S, et al. (2014) Illumina human exome genotyping array clustering and quality control. Nat Protoc 9: 2643-2662

[19] Shimajiri Y, Sanke T, Furuta H, et al. (2001) A missense mutation of Pax4 gene (R121W) is associated with type 2 diabetes in Japanese. Diabetes 50: 2864-2869

[20] Turner S, Armstrong LL, Bradford Y, et al. (2011) Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet Chapter 1: Unit1 19

[21] Purcell S, Neale B, Todd-Brown K, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575

[22] Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26: 2190-2191

[23] Li MX, Gui HS, Kwan JS, Bao SY, Sham PC (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res 40: e53
[24] Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4: 1073-1081
[25] Adzhubei IA, Schmidt S, Peshkin L, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249
[26] Plengvidhya N, Kooptiwut S, Songtawee N, et al. (2007) PAX4 mutations in Thais with maturity onset diabetes of the young. J Clin Endocrinol Metab 92: 2821-2826
[27] Smith SB, Ee HC, Conners JR, German MS (1999) Paired-homeodomain transcription factor PAX4 acts as a transcriptional repressor in early pancreatic development. Mol Cell Biol 19: 8272-8280
[28] Bernardo AS, Hay CW, Docherty K (2008) Pancreatic transcription factors and their role in the birth, life and survival of the pancreatic beta cell. Mol Cell Endocrinol 294: 1-9
[29] Blyszczuk P, Czyz J, Kania G, et al. (2003) Expression of Pax4 in embryonic stem cells promotes differentiation of nestin-positive progenitor and insulin-producing cells. Proc Natl Acad Sci U S A 100: 998-1003
[30] Sosa-Pineda B, Chowdhury K, Torres M, Oliver G, Gruss P (1997) The Pax4 gene is essential for differentiation of insulin-producing beta cells in the mammalian pancreas. Nature 386: 399-402
[31] Petersen HV, Jorgensen MC, Andersen FG, et al. (2000) Pax4 represses pancreatic glucagon gene expression. Mol Cell Biol Res Commun 3: 249-254
[32] Cho YS, Chen CH, Hu C, et al. (2012) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet 44: 67-72
[33] Ma RC, Hu C, Tam CH, et al. (2013) Genome-wide association study in a Chinese population identifies a susceptibility locus for type 2 diabetes at 7q32 near PAX4. Diabetologia 56: 1291-1305
[34] Kooptiwut S, Plengvidhya N, Chukijrungroat T, et al. (2012) Defective PAX4 R192H transcriptional repressor activities associated with maturity onset diabetes of the young and early onset-age of type 2 diabetes. J Diabetes Complications 26: 343-347
[35] Xu W, Rould MA, Jun S, Desplan C, Pabo CO (1995) Crystal structure of a paired domain-DNA complex at 2.5 A resolution reveals structural basis for Pax developmental mutations. Cell 80: 639-650
[36] Fuchsberger C, Flannick J, Teslovich TM, et al. (2016) The genetic architecture of type 2 diabetes. Nature 536: 41-47
[37] Voight BF, Scott LJ, Steinthorsdottir V, et al. (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42: 579-589
[38] Scott RA, Lagou V, Welch RP, et al. (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 44: 991-1005
[39] Steinthorsdottir V, Thorleifsson G, Reynisdottir I, et al. (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 39: 770-775
[40] Liao JC, Yang TT, Weng RR, Kuo CT, Chang CW (2015) TTBK2: a tau protein kinase beyond tau phosphorylation. Biomed Res Int 2015: 575170
[41] Alesutan I, Sopjani M, Dermaku-Sopjani M, Munoz C, Voelkl J, Lang F (2012) Upregulation of Na-coupled glucose transporter SGLT1 by Tau tubulin kinase 2. Cell Physiol Biochem 30: 458-465
[42] Cariou B, Charbonnel B (2015) Sotagliflozin as a potential treatment for type 2 diabetes mellitus. Expert Opin Investig Drugs 24: 1647-1656
[43] Hart AW, Baeza N, Apelqvist A, Edlund H (2000) Attenuation of FGF signalling in mouse beta-cells leads to diabetes. Nature 408: 864-868
[44] Woo YC, Xu A, Wang Y, Lam KS (2013) Fibroblast growth factor 21 as an emerging metabolic regulator: clinical perspectives. Clin Endocrinol (Oxf) 78: 489-496

[45] Vordenbaumen S, Braukmann A, Petermann K, et al. (2011) Casein alpha s1 is expressed by human monocytes and upregulates the production of GM-CSF via p38 MAPK. J Immunol 186: 592-601
[46] Dinarello CA, Donath MY, Mandrup-Poulsen T (2010) Role of IL-1beta in type 2 diabetes. Curr Opin Endocrinol Diabetes Obes 17: 314-321

Figure 1. Manhattan plot of discovery stage results.

Figure legend: The y-axis represents the –log10 *p*-value, and the x-axis represents the 77,468 analysed SNPs. The grey dash horizontal line indicates the exome-wide significance ($6.45x10^{-7}$). The diamond symbol indicates the exome-wide significant SNPs.

Table 1. Clinical characteristics of study participants in the discovery stage.

	Controls	Cases	*p*-value
Number	1988	3652	-
Male, %	56.4	60.9	<0.001
Age, year	58.7 ± 12.1	64.8 ± 11.8	<0.001
Fasting glucose, mmol/l	5.1 ± 0.6	7.6 ± 2.5	<0.001
Body mass index, kg/m^2	24.2 ± 3.7	25.9 ± 4.0	<0.001
Waist circumference, cm	M: 86.3 ± 8.4	M: 91.5 ± 10.2	<0.001
	F:79.1 ± 9.1	F:86.3 ± 10.9	<0.001
Coronary artery disease, %	37.2	43.8	<0.001
Hypertension, %	40.3	85.5	<0.001
Use of anti-hypertensive drug, %	31.5	83.0	<0.001
Use of lipid lowering drug, %	37.5	65.6	<0.001
Ever Smoker, %	34.4	35.9	0.251

Data as mean ± standard deviation. M: Male; F: Female; T2DM, type 2 diabetes.

Table 2. Association results of SNPs reaching exome-wide significance ($p<6.45x10^{-7}$) in the discovery stage.

Nearest gene(s)	SNP	Position	Annotation	A1	A2	MAF		OR(95%CI)	*p*-value[a]	*p*-value[b]
						Cases	Controls			
Asian-specific variant										
PAX4	rs2233580	7:127253550	p.Arg192His	T	C	0.145	0.113	1.39(1.23-1.56)	1.75×10^{-7}	7.62×10^{-6}
Established type 2 diabetes susceptibility variants										
CDKN2B-AS1/DMRTA1	rs10965250	9:22133284	intergenic	A	G	0.384	0.430	0.80(0.74-0.87)	5.93×10^{-8}	8.80×10^{-10}
KCNQ1	rs2237896	11:2858440	intronic	A	G	0.311	0.359	0.80(0.73-0.87)	1.82×10^{-7}	1.53×10^{-8}

A1: Minor allele; A2: Major allele; MAF: Minor allele frequency. The ORs are reported with respect to the minor allele.

[a]Adjusted for age, sex, PC1 and PC2. [b]Adjusted for age, sex, BMI, PC1 and PC2.

Table 3. Replication and combined association results.

			Discovery Hong Kong (3652cases VS 1988controls)		Replication Guangzhou (3537cases VS 8825controls)		Combined Hong Kong + Guangzhou (7189cases VS 10813controls)		
Gene(s)	SNP	A1	OR(95%CI)	*p*-value[a]	OR(95%CI)	*One-tailed* *p*-value[a]	Dir	OR(95%CI)	*p*-value[a]
PAX4	rs2233580	T	1.39(1.23-1.56)	1.75×10^{-7}	1.28(1.18-1.39)	$\underline{1.22 \times 10^{-9}}$	+ +	1.31(1.23-1.40)	3.74×10^{-15}
FGFR1	rs2288696	A	0.73(0.63-0.85)	2.29×10^{-5}	0.98(0.88-1.09)	0.350	– –	0.88(0.81-0.96)	4.57×10^{-3}
ANKRD55/MAP3K1	rs456867	T	0.84(0.77-0.91)	4.78×10^{-5}	0.99(0.93-1.05)	0.363	– –	0.94(0.89-0.98)	8.57×10^{-3}
TTBK2	rs56017612	C	0.72(0.61-0.84)	7.40×10^{-5}	0.90(0.80-1.02)	0.046	– –	0.83(0.75-0.92)	2.11×10^{-4}
DUSP26/UNC5D	rs4739563	T	0.80(0.72-0.90)	7.48×10^{-5}	1.00(0.93-1.08)	0.518	– +	0.93(0.87-0.99)	0.020
HCG27/HLA-C	rs3869115	C	0.81(0.73-0.90)	1.04×10^{-4}	0.99(0.92-1.07)	0.418	– –	0.93(0.87-0.99)	0.016
DAP	rs267939	A	0.79(0.69-0.89)	1.85×10^{-4}	1.01(0.92-1.10)	0.545	– +	0.92(0.86-0.99)	0.035
CSN1S1	rs10030475	T	0.85(0.78-0.92)	1.86×10^{-4}	0.93(0.87-0.99)	7.5×10^{-3}	– –	0.90(0.85-0.94)	3.28×10^{-5}
ZNF283/ZNF404	rs138993781	G	0.27(0.13-0.55)	2.93×10^{-4}	0.69(0.34-1.39)	0.148	– –	0.43(0.26-0.71)	1.03×10^{-3}
STAB1	rs740903	T	1.19(1.08-1.31)	3.02×10^{-4}	1.00(0.94-1.07)	0.450	+ +	1.06(1.01-1.12)	0.030
CARNS1	rs868167	A	0.67(0.54-0.83)	3.33×10^{-4}	1.05(0.90-1.23)	0.730	– +	0.90(0.79-1.02)	0.109

A1: Minor allele; Dir: Direction of effect. The ORs are reported with respect to the minor allele. [a]Adjusted for age and sex. One-tailed *p*-value: For effects in the same direction as in the discovery stage analysis, one-tailed *p*-values were calculated as (*p*/2); for effects in opposite direction, one-tailed *p*-values were calculated as (1 - *p*/2). SNP that survived Bonferroni correction for multiple testing in the replication analysis is underlined.

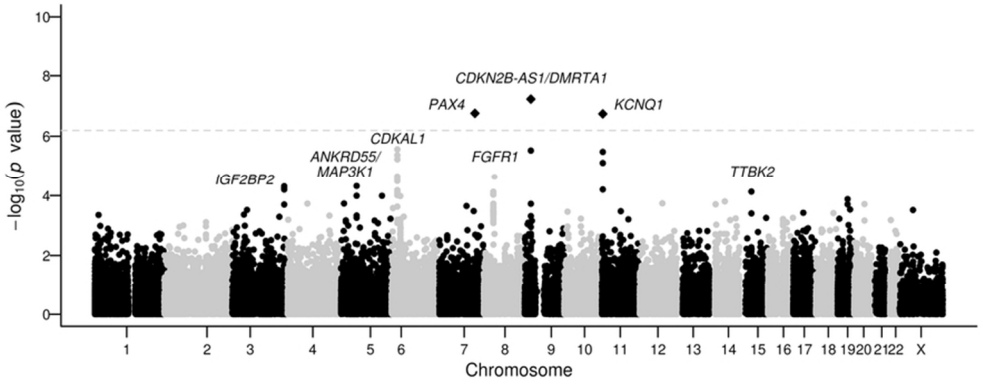

Figure 1. Manhattan plot of discovery stage results.

Figure legend: The y-axis represents the –log10 p-value, and the x-axis represents the 77,468 analysed SNPs. The grey dash horizontal line indicates the exome-wide significance (6.45×10^{-7}). The diamond symbol indicates the exome-wide significant SNPs.

76x32mm (300 x 300 DPI)

ESM Table 1. Clinical characteristics of study participants in the replication stage.

	Controls	Cases	*p-value*
Number	8825	3537	-
Male, %	22.9	22.7	0.882
Age, year	63.5 ± 7.1	65.9 ± 6.8	<0.001
Fasting glucose, mmol/l	5.0 ± 0.01	7.1 ± 2.5	<0.001
Haemoglobin A1c, %	5.8 ± 0.4	7.0 ± 1.4	<0.001
Haemoglobin A1c, mmol/mol	39.5 ± 4.5	43.3 ± 15.1	<0.001
Body mass index, kg/m^2	23.3 ± 3.2	24.9 ± 3.5	<0.001
Waist circumference, cm	M:84.6 ± 9.0 F:79.6 ± 8.7	M:87.2 ± 9.6 F:82.7 ± 9.4	<0.001 <0.001
Coronary artery disease[a], %	5.8	10.4	<0.001
Hypertension[b], %	25.8	51.0	<0.001
Use of anti-hypertensive drug, %	20.8	45.8	<0.001
Use of lipid lowering drug, %	4.2	6.1	<0.001
Ever smoker, %	18.3	18.8	0.556

Data as mean ± standard deviation. [a]Self-reported CAD. [b]Self-reported hypertension.

ESM Table 2. Association results of 34 top index SNPs with $p<5x10^{-4}$ in the discovery stage.

Gene(s)	SNP	Position	Annotation	A1	A2	MAF		OR(95%CI)	*p-value*[a]	*p-value*[b]
						Cases	Controls			
Known loci										
CDKN2B-AS1/DMRTA1	rs10965250	9:22133284	intergenic	A	G	0.384	0.430	0.80(0.74-0.87)	5.93×10^{-8}	8.80×10^{-10}
PAX4	rs2233580[c]	7:127253550	p.Arg192His	T	C	0.145	0.113	1.39(1.23-1.56)	1.75×10^{-7}	7.62×10^{-6}
KCNQ1	rs2237896	11:2858440	intronic	A	G	0.311	0.359	0.80(0.73-0.87)	1.82×10^{-7}	1.53×10^{-8}
CDKAL1	rs10440833[c]	6:20688121	intronic	A	T	0.385	0.340	1.22(1.12-1.33)	2.71×10^{-6}	3.66×10^{-7}
ANKRD55/MAP3K1	rs456867[c]	5:55811092	intronic	T	C	0.330	0.366	0.84(0.77-0.91)	4.78×10^{-5}	1.65×10^{-4}
IGF2BP2	rs11711477[c]	3:185526690	intronic	A	T	0.263	0.226	1.21(1.11-1.33)	4.88×10^{-5}	1.17×10^{-4}
CDC123/CAMK1D	rs10906115	10:12314997	intergenic	G	A	0.375	0.405	0.86(0.79-0.93)	3.46×10^{-4}	3.03×10^{-4}
HNF1B	rs7501939	17:36101156	intronic	T	C	0.238	0.204	1.19(1.08-1.31)	3.77×10^{-4}	1.04×10^{-4}
ANKRD55/MAP3K1	rs13178412	5:55831021	p.Tyr23His	G	A	0.047	0.061	0.73(0.62-0.87)	4.66×10^{-4}	2.39×10^{-4}
CDKN2B-AS1/DMRTA1	rs10965251	9:22134029	intergenic	A	G	0.072	0.087	0.77(0.66-0.89)	4.86×10^{-4}	3.18×10^{-4}
Novel loci										
FGFR1	rs2288696[c]	8:38286225	intronic	A	G	0.069	0.091	0.73(0.63-0.85)	2.29×10^{-5}	2.21×10^{-5}
TTBK2	rs56017612[c]	15:43086885	p.Thr313Ala	C	T	0.054	0.072	0.72(0.61-0.84)	7.40×10^{-5}	1.71×10^{-3}
DUSP26/UNC5D	rs4739563[c]	8:34247995	intergenic	T	C	0.148	0.174	0.80(0.72-0.90)	7.48×10^{-5}	2.35×10^{-4}
PPP2R2B/STK32A	rs6893679	5:146542753	intergenic	A	G	0.322	0.289	1.19(1.09-1.30)	1.02×10^{-4}	5.77×10^{-3}
HCG27/HLA-C	rs3869115[c]	6:31204694	intergenic	C	G	0.162	0.193	0.81(0.73-0.90)	1.04×10^{-4}	1.00×10^{-4}
SCN1B	rs67701503[c]	19:35524939	p.Ser248Arg	A	C	0.206	0.237	0.83(0.75-0.91)	1.32×10^{-4}	7.16×10^{-4}
SAMD4A/GCH1	rs8022503	14:55265828	intergenic	T	C	0.312	0.277	1.18(1.09-1.29)	1.57×10^{-4}	1.21×10^{-3}
PTPRQ	rs6539524	12:80935345	p.Phe1056Leu	C	T	0.251	0.220	1.20(1.09-1.32)	1.82×10^{-4}	2.26×10^{-3}

DAP	rs267939 [c]	5:10752315	intronic	A	G	0.099	0.123	0.79(0.69-0.89)	1.85×10^{-4}	1.88×10^{-4}
CSN1S1	rs10030475 [c]	4:70807771	p.Ala117Val	T	C	0.277	0.312	0.85(0.78-0.92)	1.86×10^{-4}	1.29×10^{-3}
KIAA1755	rs3746471	20:36841914	p.Arg1045Trp	A	G	0.430	0.467	0.86(0.79-0.93)	1.93×10^{-4}	1.68×10^{-4}
OR4E2/DAD1	rs10140810	14:22392626	intergenic	C	A	0.473	0.435	1.17(1.08-1.26)	1.96×10^{-4}	2.92×10^{-3}
ACN9/TAC1	rs7791918	7:97207509	intergenic	T	G	0.420	0.380	1.17(1.08-1.26)	2.23×10^{-4}	1.15×10^{-4}
PPP1R3G/LYRM4	rs685187	6:5106807	ncRNA	C	T	0.109	0.131	0.80(0.71-0.90)	2.62×10^{-4}	2.26×10^{-2}
ZNF283/ZNF404	rs138993781 [c]	19:44366936	intergenic	G	A	0.002	0.005	0.27(0.13-0.55)	2.93×10^{-4}	3.75×10^{-3}
STAB1	rs740903 [c]	3:52548818	p.Cys1260Cys	T	C	0.251	0.222	1.19(1.08-1.31)	3.02×10^{-4}	2.71×10^{-5}
SLC38A5	rs17281188	X:48317386	p.Met451Thr	G	A	0.077	0.095	0.73(0.62-0.87)	3.05×10^{-4}	2.02×10^{-3}
SLC13A1	rs2140516	7:122809234	p.Asn174Ser	C	T	0.337	0.374	0.86(0.79-0.93)	3.31×10^{-4}	7.66×10^{-4}
CARNS1	rs868167 [c]	11:67186271	p.Pro137Thr	A	C	0.029	0.041	0.67(0.54-0.83)	3.33×10^{-4}	2.83×10^{-4}
LOC284294/CDH7	rs531795	18:62873393	intergenic	A	C	0.475	0.500	0.86(0.80-0.94)	3.85×10^{-4}	1.61×10^{-4}
TTBK2	rs6493068	15:43170793	p.Leu8Pro	G	A	0.402	0.431	0.86(0.80-0.94)	3.96×10^{-4}	1.24×10^{-2}
ULK4	rs1795316	3:41531910	intronic	G	T	0.377	0.344	1.16(1.07-1.26)	4.32×10^{-4}	6.78×10^{-4}
PDPN	chr1:13937002 [c]	1:13937002	p.His66Tyr	T	C	0.001	0.004	0.18(0.07-0.47)	4.46×10^{-4}	8.32×10^{-4}
NR3C2/DCLK2	rs6826460	4:149605653	intergenic	G	A	0.274	0.305	0.85(0.78-0.93)	4.69×10^{-4}	9.87×10^{-4}

A1: Minor allele; A2: Major allele; A1: Minor allele; A2: Major allele. The ORs are reported with respect to the minor allele. [a]Adjusted for age, sex, PC1 and PC2. [b]Adjusted for age, sex, BMI, PC1 and PC2. SNPs with exome-wide significance ($p<6.45 \times 10^{-7}$) are underlined. [c]SNPs followed-up in the replication analysis.

ESM Table 3. Distribution of mean age of diagnosis by *PAX4* rs2233580 genotypes in the current study and 3 independent cohorts

Study	No. of Cases	Mean age of diagnosis ± S.D (years) CC	CT	TT	z-score	*p-value*
Hong Kong Chinese[a]: HKU-TRS	3652	2666 54.37±12.63	910 52.73±13.02	76 52.37±12.69	-3.431	- 6.01×10^{-4}
Singapore Chinese[b]: Singapore Diabetes Cohort Study and Singapore Prospective Study Programme	560	417 57.12±12.55	130 56.47±13.03	13 52.40±12.18	-1.195	- 0.232
Hong Kong Chinese[b]: Hong Kong Diabetes Registry (CUHK)	489	314 36.41±9.49	153 36.80±10.13	22 32.14±6.58	0.931	- 0.351
Korean[b]: Seoul National University Hospital Diabetes Case Control Study (SNUH)	570	458 50.38±9.67	98 51.61±11.52	14 51.20±10.90	0.742	- 0.458
Combined	5271	-	-	-	-2.717	0.007[c]

S.D, standard deviation. [a] Current study. [b] Reported in reference [1]. [c] *p*-value from a sample-size weighted meta-analysis by combining unadjusted *p*-values across studies using METAL [2] (I^2=58.5, $p_{heterogeneity}$=0.065).

Reference

1. Fuchsberger C, Flannick J, Teslovich TM, et al. (2016) The genetic architecture of type 2 diabetes. Nature 536: 41-47

2. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26: 2190-2191